Bernoulli **27**(3), 2021, 1467–1495
https://doi.org/10.3150/20-BEJ1280

Inference without compatibility: Using exponential weighting for inference on a parameter of a linear model

MICHAEL LAW* and YA'ACOV RITOV†

Department of Statistics, University of Michigan, Ann Arbor, USA.
E-mail: *mmylaw@umich.edu; †yritov@umich.edu*

We consider hypotheses testing problems for three parameters in high-dimensional linear models with minimal sparsity assumptions of their type but without any compatibility conditions. Under this framework, we construct the first $\sqrt{n}$-consistent estimators for low-dimensional coefficients, the signal strength, and the noise level. We support our results using numerical simulations and provide comparisons with other estimators.

Keywords: Lasso; compatibility condition; exponential weighting; inference

1. Introduction

In the past decade, there has been substantial interest in high-dimensional linear models, particularly following the work of Tibshirani [25]. However, it was not until the past few years that there have been methods to construct confidence intervals and p-values for particular covariates in the model. Consider a high-dimensional partially linear model

$$Y = X\beta + \mu + \varepsilon, \tag{1.1}$$

with $X \in \mathbb{R}^{n\times q}$, and $Y, \mu, \varepsilon \in \mathbb{R}^n$. In addition, we also observe covariates $Z \in \mathbb{R}^{n\times p}$ such that $\mu \approx Z\gamma$ for some sparse vector $\gamma \in \mathbb{R}^p$ (see Section 1.2 for details). The vector μ represents some underlying random nuisance parameters in the model that affect the response Y; the covariates Z allow us to control for these confounding factors. Regarding the size of each matrix, we assume that $q < n$ is fixed but $p > n$ is high-dimensional. Our goal is to construct a confidence region for the entire vector $\beta \in \mathbb{R}^q$.

In recent years, there have been mainly two approaches to constructing confidence intervals in high-dimensional linear models. On one hand, authors like Lee *et al.* [18] construct conditional confidence intervals for β given that β was selected by a procedure, such as the lasso. Simultaneously, there has been work to construct unconditional confidence intervals for β, where X is the a priori selected covariate of interest, such as Javanmard and Montanari [15], van de Geer *et al.* [27], and Zhang and Zhang [30]; the latter is also our focus. To avoid digressions, we will not elaborate on the former. A review of many of the current methods is available in Dezeure *et al.* [7]. Much of the existing literature relies on using a version of the de-sparsified lasso introduced simultaneously by Javanmard and Montanari [15], van de Geer *et al.* [27], and Zhang and Zhang [30]. The idea behind the existing approaches is to invert the KKT conditions of the lasso and perform nodewise lasso to approximate the inverse covariance matrix of the design, which attempts to correct the bias introduced by the lasso.

Since the lasso forms the basis for the procedure, certain assumptions must be made in order to ensure that the lasso enjoys the nice theoretical properties that have been developed over the past two decades. The paper by van de Geer and Bühlmann [26] provides an overview of various assumptions

that have been used to prove oracle inequalities for the lasso. These assumptions are a consequence of the fact the lasso is used rather than being needed for the statistical problem. In particular, for confidence intervals, van de Geer *et al.* [27] assume that the compatibility condition holds for the Gram matrix, which is the weakest assumption from van de Geer and Bühlmann [26], and is essentially a necessary assumption for the lasso to enjoy the fast rate (cf. Bellec [1]). To quote the popular book by Bühlmann and van de Geer [4], "In fact, a compatibility condition is nothing else than simply an assumption that makes our proof go through." However, this raises an important question on necessity: Is the compatibility condition necessary for constructing confidence intervals in high-dimensions?

The main contribution of this paper is proving that the compatibility condition or any of its variants is indeed not necessary for the statistical problem. To this end, we provide an estimator that does not require the compatibility condition but still attains the semi-parametric efficiency bound. Our assumption regarding sparsity is slightly stronger than the minimax rate required by Javanmard and Montanari [16] since we allow a broader class of designs. In particular, we show that, in the absence of compatibility, the rate established by Javanmard and Montanari [16] is unattainable and a stronger sparsity assumption is required.

To help clarify the connection between our notion of partially linear model and the high-dimensional linear models of the aforementioned works, we note that our model is many times written as a linear model $y = x^\mathsf{T}\beta + z^\mathsf{T}\gamma + \varepsilon$, reserving the notion of partially linear model to $y = x^\mathsf{T}\beta + \mu(t) + \varepsilon$ for some unknown smooth function $\mu(\cdot)$. We use the partially linear model terminology to emphasize that (i) $z^\mathsf{T}\gamma$ is only an approximation, and (ii) z is a high-dimensional nuisance parameter, which plays the role of the nonparametric component of a semi-parametric model. For more details, see Remark 1.1 below.

There is also the recent work of Chernozhukov *et al.* [6], who consider the general problem of conducting inference on low-dimensional parameters with high-dimensional nuisance parameters. One application of their general theory is for high-dimensional partially linear models, which is also our problem of interest. A further discussion of their procedure is given in Remark 2.4 below.

As a consequence of our estimation procedure for β, we are able to construct a $\sqrt{n}$-consistent estimator of the signal strength and the noise variance, which we denote by σ_μ^2 and σ_ε^2 respectively, also without the compatibility condition. The paper by Reid, Tibshirani and Friedman [22] provides a nice overview of different proposals for estimation of σ_ε^2 using the lasso. An early work in this direction is Fan, Guo and Hao [9], who construct asymptotic confidence intervals for σ_ε^2 under a sure screening property of the covariates; in the setting of the lasso, this requires a β-min condition. Dicker [8] considers a similar problem of variance estimation using moment estimators that do not require sparsity of the underlying signal. However, he does not consider the ultra high-dimensional setting nor the problem of inference. Later, Janson, Barber and Candès [14] considered inference on the signal to noise ratio but the theory developed only applies to Gaussian designs. For the problem of inference for σ_μ^2, the work most similar with ours is Cai and Guo [5], who consider a more general problem in the semi-supervised setting, but their results for the supervised framework require minimal non-zero eigenvalues on the covariance matrix. To this end, we construct estimators that attain asymptotic variances equal to that of the efficient estimator in low-dimensions.

For both problems, our approach involves using exponential weighting to aggregate over all models of a particular size. Prima facie, this is a computationally hard problem but can be well approximated in practice. To this end, we propose an algorithm inspired by Rigollet and Tsybakov [23].

1.1. Organization of the paper

We will end the current section with the notation that will be used throughout the paper. In Section 2, we discuss the problem of conducting inference for low-dimensional β in the presence of a high-

dimensional nuisance vector μ. The setting of univariate β is considered separately in Section 2.1 to motivate the general multivariate procedure of Section 2.3. We take a slight detour in Section 2.2 to consider inference when the errors are correlated. The section ends with a discussion on the necessity of the sparsity assumption in Section 2.4. Then, in Section 3.1 and Section 3.2, we consider the problems of inference for σ_μ^2 and σ_ε^2, respectively. In Section 4, we provide an overview of the computation of the estimators, which we apply in Section 5 for numerical simulations. The proofs for Sections 2.1 and 2.4 are provided in Section 6. Additional simulation tables and the proofs for the remaining results are available in the Supplement [17].

1.2. General notation and definitions

Throughout, all of our variables (except β) have a dependence on n, but when it should not cause confusion, this dependence will be suppressed. For a general vector a and a matrix A, a_j will denote the j'th entry of a, A_j the j'th column of A, and $A^{(j)}$ the j'th row of A. Then, $\|a\|$ will denote the standard Euclidean norm, with the dimension of the space being implicit from the vector, $\|a\|_1$ the L_1-norm, and $\|a\|_0$ the L_0-norm. Furthermore, $\|A\|$ will denote the operator norm and $\|A\|_{\mathrm{HS}}$ the Hilbert–Schmidt norm. If A is square, A^{-1} is to be interpreted in a generalized sense whenever the matrix A is rank deficient.

Before defining weak sparsity, we will need to introduce some notation. For $u \in \mathbb{N}$, $\mathcal{M}_u$ will denote the collection of all models of Z of size u. That is,

$$\mathcal{M}_u \triangleq \big\{m \subseteq \{1, \dots, p\} : |m| = u\big\}.$$

Then, for each $m \in \mathcal{M}_u$, Z_m will denote the $n \times u$ sub-matrix of Z corresponding to the columns indexed by m. Moreover, P_m will denote the projection onto the column space of Z_m and $P_m^\perp$ the projection onto the orthogonal complement. We can now state the definition of weak sparsity.

Definition 1.1. A sequence of vectors μ is said to satisfy the *weak sparsity property relative to* Z with sparsity s at rate k if the set

$$\mathcal{S}_\mu \triangleq \big\{m \in \mathcal{M}_s : \big\|P_m^\perp \mu\big\|^2 = o(k)\big\}$$

is non-empty. A set $S \in \mathcal{S}_\mu$ is said to be a *weakly sparse set* for the vector μ.

If the sequence of vectors μ_n is random, then it satisfies the *weak sparsity property relative to* Z *in probability* with sparsity s at rate k if the set

$$\mathcal{S}_\mu = \big\{m \in \mathcal{M}_s : \big\|P_m^\perp \mu\big\|^2 = o_{\mathbb{P}}(k)\big\}$$

is nonempty. A set $S \in \mathcal{S}_\mu$ is said to be a *weakly sparse set in probability* for the vector μ.

Remark 1.1. There are two distinctions to be made, between *strong* and *weak* sparsity on one hand, and between *weak sparsity* and *weak sparsity in probability* on the other. The following examples may help to clarify these notions.

First, suppose that $\mu = Z\gamma$ for a sparse vector $\gamma \in \mathbb{R}^p$ with support S. We refer to this case as *strong sparsity* and is the commonly assumed setting in high-dimensional linear models (for example, see van de Geer *et al.* [27]). Since $\|P_S^\perp \mu\|^2 = 0$, strong sparsity implies weak sparsity.

Second, consider a smooth function $\mu(\cdot) : \mathbb{R} \to \mathbb{R}$. This corresponds to a standard partially linear model, where $\mu(t)$ depends on time. Let Z be a dictionary of basis functions, say, the harmonic or

wavelet basis. Then, μ may be well approximated by a linear combination of a few basis functions, with the remainder converging to zero, and weak sparsity holds.

Third, in random designs, on a set with small probability, μ may not be well approximated by any members of $\mathcal{M}_s$, while it is well approximated on the complement. This case is referred to as weak sparsity in probability.

In general, if $\mathcal{S}_\mu$ is non-empty, then we may let $\gamma = (Z_S^\mathsf{T} Z_S)^{-1} Z_S \mu$ for any $S \in \mathcal{S}_\mu$. Depending on context, we will either view γ as a vector in $\mathbb{R}^p$ or $\mathbb{R}^s$.

Finally, similar to other works on de-biased inference, we will consider sub-Gaussian errors, which is defined below.

Definition 1.2. A mean zero random vector $\xi \in \mathbb{R}^n$ is said to be *sub-Gaussian* with parameter K if

$$\mathbb{E} \exp\left(\lambda^\mathsf{T} \xi\right) \leq \exp\left(\frac{K^2 \|\lambda\|^2}{2}\right)$$

for all vectors $\lambda \in \mathbb{R}^n$.

2. Inference for β

In this section, we consider the main problem of constructing confidence regions for β. The model that we consider is given in equation (1.1), which we reproduce below for convenience,

$$Y = X\beta + \mu + \varepsilon. \tag{2.1}$$

We will write $\sigma_\varepsilon^2 \triangleq \mathrm{Var}(\varepsilon_1)$. For this section, we will assume that μ satisfies the weak sparsity property relative to Z at rate $\sqrt{n}$, but the results still hold if we assume the weak sparsity property in probability.

2.1. The special case: $q = 1$

In this sub-section, we will assume throughout that $q = 1$. In addition to the partially linear model given in equation (2.1), we also assume that X satisfies a partially linear model, denoted by

$$X = \nu + \eta, \tag{2.2}$$

where ν satisfies the weak sparsity property relative to Z at rate $\sqrt{n}$. We allow the weakly sparse set for ν to be different from that of μ. We will also assume that η is a sub-Gaussian vector with variance $\sigma_\eta^2 \triangleq \mathrm{Var}(\eta_1)$. The sub-Gaussianity assumption is needed to ensure that the empirical estimate of the norm squared residuals approximates the expectation well enough. Note that this structural assumption is implied if we assume the design is Gaussian, which is a common assumption in the literature. In this setting, the distribution of η will be Gaussian and the weak sparsity property would be satisfied if the inverse covariance matrix is row sparse. By direct substitution, it follows that

$$Y = \nu\beta + \mu + \eta\beta + \varepsilon.$$

Since μ and ν both satisfy the weak sparsity property relative to Z at rate $\sqrt{n}$, the vector $\nu\beta + \mu$ also satisfies the weak sparsity property relative to Z at rate $\sqrt{n}$.

To motivate our procedure, we will assume temporarily that the models are in fact low-dimensional linear models. That is, suppose there are sets S_δ and S_γ such that $\nu = Z_{S_\delta}\delta$ and $\mu = Z_{S_\gamma}\gamma$ for sparse vectors δ and γ. Moreover, assume that the set $S \triangleq S_\delta \cup S_\gamma$ is known and $\varepsilon \sim \mathcal{N}_n(0_n, \sigma_\varepsilon^2 I_n)$. Thus, we are temporarily assuming the low-dimensional linear models

$$Y = X\beta + Z_{S_\gamma}\gamma + \varepsilon = Z_S\theta + \eta\beta + \varepsilon,$$
$$X = Z_{S_\delta}\delta + \eta,$$

where $\theta = \delta\beta + \gamma$. Then, by the Gauss-Markov theorem, it is known that the efficient estimator in this low-dimensional problem is given by least-squares, which may be framed as the following three stage procedure:

1. Regress Y on Z_S using least-squares to obtain the fitted values $\hat{Y}$.
2. Regress X on Z_S using least-squares to obtain the fitted values $\hat{X}$.
3. Regress the residuals $Y - \hat{Y}$ on the the residuals $X - \hat{X}$ using least-squares to obtain the least-squares estimator $\hat{\beta}_{\mathrm{LS}}$.

In the high-dimensional setting, the first two stages can no longer be achieved using the classical least-squares approach. However, since we are only interested in the fitted values $\hat{Y}$ and $\hat{X}$, this suggests using a high-dimensional prediction procedure to obtain the fitted values, and then applying low-dimensional least-squares on the residuals in the third stage. The high-dimensional procedure that we will adopt is the exponential weights of Leung and Barron [19], which has the salient feature of prediction consistency under very mild assumptions on the design.

Before defining our estimators, we will state all of our assumptions.

(A1) The means μ and ν have squared norms that are $\mathcal{O}_{\mathbb{P}}(n)$.

(A2) The entries of η and ε are mutually independent and also independent of Z. Moreover, the entries of η and ε are each identically distributed sub-Gaussians with parameters K_η and K_ε, respectively.

(A3) The means μ, ν, and $\nu\beta + \mu$ are weakly sparse relative to Z with sparsities s_γ, s_δ, and s_θ, respectively at rate $\sqrt{n}$. Furthermore, it is assumed that the statistician knows sequences u_γ, u_δ and u_θ with $u_\gamma \geq s_\gamma$, $u_\delta \geq s_\delta$, and $u_\theta \geq s_\theta$ for n sufficiently large and $\max(u_\gamma, u_\delta, u_\theta) = o(\sqrt{n}/\log(p))$.

Condition (A1) avoids the trivial situations where the signal to noise ratios, $\|\mu\|^2/n\sigma_\varepsilon^2$ and $\|\nu\|^2/n\sigma_\eta^2$, approach either zero or infinity. If either $\|\mu\|^2/n\sigma_\varepsilon^2 \to 0$ or $\|\nu\|^2/n\sigma_\eta^2 \to \infty$, the information in estimating β will approach zero as $\sigma_\varepsilon^2/\sigma_\eta^2 \to \infty$. Conversely, if either $\|\mu\|^2/n\sigma_\varepsilon^2 \to \infty$ or $\|\nu\|^2/n\sigma_\eta^2 \to 0$, then the asymptotic distribution will be degenerate, even in the low-dimensional problem.

Now, we may define two sets of exponential weights, $w_{m,Y}$ and $w_{m,X}$, to estimate $\hat{Y}$ and $\hat{X}$, respectively. Let

$$w_{m,Y} \triangleq \frac{\exp(-\frac{1}{\alpha_Y}\|P_m^\perp Y\|^2)}{\sum_{k\in\mathcal{M}_{u_\theta}}\exp(-\frac{1}{\alpha_Y}\|P_k^\perp Y\|^2)},$$

with $\alpha_Y > 4K_\varepsilon^2$.

Remark 2.1. The exponential weights defined above do not subtract off the rank of the projection in the exponent as in Leung and Barron [19] since we only consider models of size u_θ; the rank will cancel from the numerator and the denominator.

Now, let $\hat{\theta}_m \triangleq (Z_m^\mathsf{T} Z_m)^{-1} Z_m^\mathsf{T} Y$ be the least-squares estimator for θ using the covariates Z_m. We will identify $\hat{\theta}_m$ with a vector in $\mathbb{R}^p$, with the support of $\hat{\theta}_m$ being indexed by m. Then, we may estimate θ by

$$\hat{\theta}_{\mathrm{EW}} \triangleq \sum_{m \in \mathcal{M}_{u_\theta}} w_{m,Y} \hat{\theta}_m,$$

with the prediction $\hat{Y}$ given by $\hat{Y} = Z\hat{\theta}_{\mathrm{EW}}$. Similarly, we will define

$$w_{m,X} \triangleq \frac{\exp(-\frac{1}{\alpha_X} \| P_m^\perp X \|^2)}{\sum_{k \in \mathcal{M}_{u_\delta}} \exp(-\frac{1}{\alpha_X} \| P_k^\perp X \|^2)},$$

with $\alpha_X > 4K_\eta^2$. Letting $\hat{\delta}_m$ denote the least-squares estimator of δ using the covariates Z_m and identifying it with a vector in $\mathbb{R}^p$, we may define

$$\hat{\delta}_{\mathrm{EW}} \triangleq \sum_{m \in \mathcal{M}_{u_\delta}} w_{m,X} \hat{\delta}_m.$$

Then, the fitted values of X will be $\hat{X} = Z\hat{\delta}_{\mathrm{EW}}$. Finally, for the last stage, the regression of $Y - Z\hat{\theta}_{\mathrm{EW}}$ on $X - Z\hat{\delta}_{\mathrm{EW}}$ will be given by

$$\hat{\beta}_{\mathrm{EW}} \triangleq \frac{(X - Z\hat{\delta}_{\mathrm{EW}})^\mathsf{T} (Y - Z\hat{\theta}_{\mathrm{EW}})}{\| X - Z\hat{\delta}_{\mathrm{EW}} \|^2}.$$

Before stating our main result, we will state a proposition regarding exponential weighting with sub-Gaussian errors.

Proposition 2.1. *Consider a high-dimensional linear model given by*

$$Y = \mu + \xi,$$

for ξ sub-Gaussian with parameter K_ξ. Assume that μ is weakly sparse relative to Z with sparsity s and that $\limsup_{n\to\infty} \|\mu\|^2 = \mathcal{O}(n)$. Assume further that the chosen sequence of sparsities $u \geq s$ satisfy $u = o(n^\tau / \log(p))$ for $\tau > 0$ fixed. Letting $\hat{\gamma}_m$ denote the least-squares estimator for γ using the covariates Z_m, define the exponential weights as

$$w_m \triangleq \frac{\exp(-\frac{1}{\alpha} \| P_m^\perp Y \|^2)}{\sum_{k \in \mathcal{M}_u} \exp(-\frac{1}{\alpha} \| P_k^\perp Y \|^2)},$$

with $\alpha > 4K_\xi^2$. Then,

$$\mathbb{E} \left\| \sum_{m \in \mathcal{M}_u} w_m Z \hat{\gamma}_m - \mu \right\|^2 = o\big(n^\tau\big).$$

Remark 2.2. We would like to remark that the choice of α is consistent with Leung and Barron [19]. In particular, when $\xi \sim \mathcal{N}_n(0_n, \sigma_\xi^2 I_n)$, the sub-Gaussian parameter is $K^2 = \sigma_\xi^2$, which gives the requirement that $\alpha > 4\sigma_\xi^2$. In this setting, we would like to emphasize that the required value of α is not

consistent with a simple Bayesian interpretation since the Bayes procedure requires a leading constant of 2, as shown by Leung and Barron [19]. However, one of the referees pointed out that Grünwald and van Ommen [12] show a way of explaining this in a Bayesian way in some extended models.

Remark 2.3. The assumption that $\limsup_{n\to\infty} \|\mu\|^2 = \mathcal{O}(n)$ can be relaxed to hold in probability by weakening the conclusion to hold in probability rather than expectation (cf. Corollary 6.4).

For the remainder of the paper, we will only consider the setting where $\tau = 1/2$. As an immediate corollary, we have the following.

Corollary 2.2. *Consider the models given in equations* (2.1) *and* (2.2) *with* $q = 1$. *Under assumptions* (A1)–(A3),

$$\|\nu\beta + \mu - Z\hat{\theta}_{\mathrm{EW}}\|^2 = o_{\mathbb{P}}(\sqrt{n}),$$
$$\|\nu - Z\hat{\delta}_{\mathrm{EW}}\|^2 = o_{\mathbb{P}}(\sqrt{n}).$$

Finally, we can state the main result for $\hat{\beta}_{\mathrm{EW}}$.

Theorem 2.3. *Consider the models given in equations* (2.1) *and* (2.2) *with* $q = 1$. *Under assumptions* (A1)–(A3),

$$\sqrt{n}(\hat{\beta}_{\mathrm{EW}} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right).$$

We would like to note that $\hat{\beta}_{\mathrm{EW}}$ attains the information bound for estimating β (cf. Example 2.4.5 of Bickel *et al.* [2] and Section 2.3.3 of van de Geer *et al.* [27]). Moreover, the convergence of $\hat{\beta}_{\mathrm{EW}}$ is actually uniform. Consider the following parameter space

$$\mathcal{B} \triangleq \left\{(\beta, \sigma_\eta^2, \sigma_\varepsilon^2, K_\eta, K_\varepsilon) : \beta \in \mathbb{R}, \sigma_\eta^2 > 0, \sigma_\varepsilon^2 > 0, K_\eta > 0, K_\varepsilon > 0\right\}.$$

This induces a set of probability measures $(\mathscr{P}_\vartheta)_{\vartheta\in\mathcal{B}}$. Then, we have the following corollary.

Corollary 2.4. *Let* $\mathscr{K}$ *be a compact set of* $(\mathscr{P}_\vartheta)_{\vartheta\in\mathcal{B}}$ *with respect to variational distance. Under the setup of Theorem* 2.3,

$$\sqrt{n}(\hat{\beta}_{\mathrm{EW}} - \beta) = A + B,$$

where

$$A \sim \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right),$$
$$|B| = o_{\mathbb{P}}(1),$$

uniformly for $\vartheta \in \mathscr{K}$.

Corollary 2.4 asserts that $\hat{\beta}_{\mathrm{EW}}$ is uniformly Gaussian regular. Like Theorem 2.3 of van de Geer *et al.* [27], the estimator $\hat{\beta}_{\mathrm{EW}}$ is regular on one-dimensional parametric sub-models of (14) of van de Geer *et al.* [27] and attains asymptotic semi-parametric efficiency. The main difference is replacing the assumption of compatibility of the design with the sparsity assumption (A3).

Remark 2.4. The estimator, $\hat{\beta}_{\mathrm{EW}}$, at first glance seems similar to the double/de-biased estimator of Chernozhukov *et al.* [6] by considering exponential weighting as the estimation procedure for the propensity function. However, the primary difference is that we do not rely on cross fitting to estimate the conditional mean of X and Y given the covariates Z. Therefore, $\hat{\beta}_{\mathrm{EW}}$ does not fall within the general framework of Chernozhukov *et al.* [6] since we are using exponential weighting to solve in the in-sample prediction problem.

To construct confidence intervals, we will need to estimate both σ_ε^2 and σ_η^2. We will defer explicitly defining estimators for the variance until Section 3.2 but let $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\eta^2$ be any of the three estimators proposed by Theorem 3.4 for estimating variance. Then, an asymptotic $(1-\alpha)$ confidence interval for β is given by

$$\left(\hat{\beta}_{\mathrm{EW}} - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\eta^2 n}}, \hat{\beta}_{\mathrm{EW}} + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\eta^2 n}}\right),$$

where $z_{\alpha/2}$ denotes the $\alpha/2$ upper quantile of the standard Gaussian distribution.

2.2. Correlated Gaussian errors

In this sub-section, we take a slight detour away from classical high-dimensional partially linear models and consider the setting where the errors, ε, are Gaussian but not necessarily independent and identically distributed. The goal is to conduct inference on β, but, for simplicity, we will only consider the setting where $q = 1$. This model arises naturally if the model was a linear mixed model given by

$$Y = X\beta + \mu + W\zeta + \xi,$$

where ζ are Gaussian random effects and ξ is independent Gaussian noise. Bradic, Claeskens and Gueuning [3] and Li, Cai and Li [20] consider more general problems of testing fixed effects in high-dimensional linear mixed models, whereas we simply view the problem as a linear model with correlated noise. Even when the errors are correlated, $\hat{\beta}_{\mathrm{EW}}$ still has a Gaussian limit under proper rescaling. Before stating the theorem, we will slightly modify assumption (A2) to the setting where ε is correlated:

(A2*) The entries of $\eta \sim \mathcal{N}_n(0_n, \sigma_\eta^2 I_n)$ are independent of Z and ε. The vector $\varepsilon \sim \mathcal{N}_n(0_n, \Sigma_\varepsilon)$ is independent of Z with $\|\Sigma_\varepsilon\| = \mathcal{O}(1)$ and $\mathrm{tr}(\Sigma_\varepsilon)/n \to \bar{d} > 0$.

Now, we may state the main result for $\hat{\beta}_{\mathrm{EW}}$ under correlation.

Theorem 2.5. *Consider the models given in equations* (2.1) *and* (2.2) *with* $q = 1$. *Under assumptions* (A1), (A2*), *and* (A3),

$$\sqrt{n}(\hat{\beta}_{\mathrm{EW}} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\bar{d}}{\sigma_\eta^2}\right).$$

Again, we will defer defining an estimator for $\bar{d}$ and σ_η^2 until Section 3.2, in particular Corollary 3.5. Similar to the previous section, we may now construct confidence intervals for β under this setting of correlation.

2.3. The general case: $q > 1$

In the general setting where $q > 1$, we may still rely on the perspective of high-dimensional prediction. Analogous to Section 2.1, we will assume that each column of X satisfies a partially linear model. That is, there exist matrices $N, H \in \mathbb{R}^{n\times q}$ (read, capital N and capital H, respectively) such that each column of X satisfies $X_j = N_j + H_j$, where N_j satisfies the weak sparsity property relative to Z at rate $\sqrt{n}$ for each $1 \le j \le q$. The weakly sparse set for each N_j may be different but the sparsity rate is uniformly $\sqrt{n}$. In matrix form, we have that

$$X = N + H. \tag{2.3}$$

Since q is fixed and μ and each N_j satisfy the weak sparsity property relative to Z at rate $\sqrt{n}$, the vector $N\beta + \mu$ also satisfies the weak sparsity property relative to Z at rate $\sqrt{n}$. Moreover, H is assumed to be sub-Gaussian with the covariance matrix of each row of H given by $\Sigma_H \triangleq \mathrm{Var}(H^{(1)})$.

Then, for $1 \le j \le q$, we may let $\hat{\delta}_{\mathrm{EW},j}$ denote the analogue of $\hat{\delta}_{\mathrm{EW}}$ for regressing X_j on Z and estimate X_j by $Z\hat{\delta}_{\mathrm{EW},j}$. Let $\hat{\Delta}_{\mathrm{EW}} \in \mathbb{R}^{p\times q}$ denote the matrix with columns given by $\hat{\delta}_{\mathrm{EW},j}$ for $1 \le j \le q$. Then, the multidimensional analogue of $\hat{\beta}_{\mathrm{EW}}$ from Section 2.1 is given by

$$\hat{\beta}_{\mathrm{EW}} \triangleq \big((X - Z\hat{\Delta}_{\mathrm{EW}})^{\mathsf{T}}(X - Z\hat{\Delta}_{\mathrm{EW}})\big)^{-1}(X - Z\hat{\Delta}_{\mathrm{EW}})^{\mathsf{T}}(Y - Z\hat{\theta}_{\mathrm{EW}}).$$

We would like to emphasize that the definition here is identical to that given in Section 2.1 when $q = 1$.

Then, we will make the following assumptions.

(B1) The mean vectors μ and N_j for $1 \le j \le q$ have squared norms that are uniformly $\mathcal{O}_{\mathbb{P}}(n)$.

(B2) The rows of H and the entries of ε are independent and also independent of Z. Moreover, the entries of the rows of H and the entries of ε are each identically distributed sub-Gaussian with parameters $K_{\eta,j}$ and K_ε respectively. Furthermore, Σ_H is an invertible matrix.

(B3) All the mean vectors μ, N_j for $1 \le j \le q$, and $N\beta + \mu$ are weakly sparse relative to Z with sparsities s_γ, $s_{\delta,j}$ for $1 \le j \le q$, and s_θ, respectively at rate $\sqrt{n}$. Furthermore, it is assumed that the statistician knows sequences u_γ, $u_{\delta,j}$, and u_θ with $u_\gamma \ge s_\gamma$, $u_{\delta,j} \ge s_{\delta,j}$ for $1 \le j \le q$ and $u_\theta \ge s_\theta$ for n sufficiently large and $\max(u_\gamma, \max_{1\le j\le q}(u_{\delta,j}), u_\theta) = o(\sqrt{n}/\log(p))$.

We can now state the asymptotic distribution for $\hat{\beta}_{\mathrm{EW}}$.

Theorem 2.6. *Consider the models given in equations* (2.1) *and* (2.3). *Under assumptions* (B1)–(B3),

$$\sqrt{n}(\hat{\beta}_{\mathrm{EW}} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}_q\big(0_q, \sigma_\varepsilon^2 \Sigma_H^{-1}\big).$$

Similar to before, to construct confidence regions, we will need to estimate Σ_H. Therefore, we will consider

$$\hat{\Sigma}_H \triangleq \frac{1}{n}(X - Z\hat{\Delta}_{\mathrm{EW}})^{\mathsf{T}}(X - Z\hat{\Delta}_{\mathrm{EW}}).$$

This leads to the following proposition.

Proposition 2.7. *Consider the models given in equations* (2.1) *and* (2.3). *Under assumptions* (B1), (B2), *and* (B3),

$$\hat{\Sigma}_H \xrightarrow{\mathbb{P}} \Sigma_H.$$

Then, an asymptotic $(1-\alpha)$ confidence region for β is given by

$$\left\{\beta \in \mathbb{R}^q : \frac{n}{\hat{\sigma}_\varepsilon^2}(\hat{\beta}_{\mathrm{EW}} - \beta)^\mathsf{T} \hat{\Sigma}_H (\hat{\beta}_{\mathrm{EW}} - \beta) \le \chi^2_{q,\alpha}\right\},$$

where $\chi^2_{q,\alpha}$ denotes the α upper quantile of a χ^2_q random variable.

2.4. Necessity of sparsity assumption

In Section 2.1, it was assumed that both μ and ν are weakly sparse with sparsity s_γ and s_δ respectively at rate $\sqrt{n}$ in order for $\hat{\beta}_{\mathrm{EW}}$ to have an asymptotic Gaussian distribution. For simplicity, in the ensuing discussion, we will only consider the case where $q = 1$, there exists an $S \in \mathcal{S}_\mu$ such that $\|P_S^\perp \mu\|^2 = 0$, and the design (X, Z) is fully Gaussian with population covariance matrix Σ. That is, $\Sigma = \mathrm{Var}((X_1, (Z^{(1)})^\mathsf{T})^\mathsf{T})$. We will write $\Sigma_{Z,Z}$ to denote the $p \times p$ sub-block of Σ corresponding to Z. Letting $\Omega = \Sigma^{-1}$, it follows that

$$s_\delta = \left|\{1 \le j \le p : \Omega_{1,j} \ne 0\}\right|,$$

which is equivalent to s_Ω from Javanmard and Montanari [16]. Compared to the de-biased lasso, Javanmard and Montanari [16] showed that, if $s_\gamma = o(n/\log^2(p))$ and $\min(s_\gamma, s_\delta) = o(\sqrt{n}/\log(p))$, then the de-biased lasso has an asymptotic Gaussian distribution. However, $\hat{\beta}_{\mathrm{EW}}$ is a valid estimator on a larger class of designs, in particular incompatible designs, and Theorem 2.8 below formalizes this trade-off between sparsity and compatibility. Before stating the theorem, we will need to introduce a bit of notation regarding our parameter space Θ, which is defined as

$$\Theta(s_\gamma, s_\delta) \triangleq \left\{\vartheta = \left(\beta, \gamma, \delta, \Sigma_{Z,Z}, \sigma_\eta^2, \sigma_\varepsilon^2\right) : \|\gamma\|_0 \le s_\gamma, \|\delta\|_0 \le s_\delta, \right.$$
$$\left. \max\left(\gamma^\mathsf{T}\Sigma_{Z,Z}\gamma, \delta^\mathsf{T}\Sigma_{Z,Z}\delta, \sigma_\eta^2, \sigma_\varepsilon^2\right) = \mathcal{O}(1)\right\}.$$

Theorem 2.8. *For $\vartheta \in \Theta(s_\gamma, s_\delta)$, consider the following model*

$$\begin{aligned} Z^{(1)}, \ldots, Z^{(n)} &\overset{i.i.d.}{\sim} \mathcal{N}_p(0_p, \Sigma_{Z,Z}), \\ \varepsilon &\sim \mathcal{N}_n\left(0_n, \sigma_\varepsilon^2 I_n\right), \\ \eta &\sim \mathcal{N}_n\left(0_n, \sigma_\eta^2 I_n\right), \\ Y &= X\beta + Z\gamma + \varepsilon, \\ X &= Z\delta + \eta. \end{aligned}$$

Assume that either $s_\gamma = o(\sqrt{n}/\log(p))$ or $s_\delta = o(\sqrt{n}/\log(p))$. If there exists a $\sqrt{n}$-consistent estimator of β for all $\vartheta \in \Theta(s_\gamma, s_\delta)$, then both $s_\gamma = \mathcal{O}(\sqrt{n}/\log(p))$ and $s_\delta = \mathcal{O}(\sqrt{n}/\log(p))$.

In light of the results of Javanmard and Montanari [16], to construct a $\sqrt{n}$-consistent estimator of β, it must be the case that either $s_\gamma = o(\sqrt{n}/\log(p))$ or $s_\delta = o(\sqrt{n}/\log(p))$. The previous theorem implies that the other sparsity must satisfy $\mathcal{O}(\sqrt{n}/\log(p))$. Assumption (A3) is only mildly stronger, requiring $\max(s_\gamma, s_\delta) = o(\sqrt{n}/\log(p))$.

Corollary 2.9. *For $\vartheta \in \Theta(s_\gamma, s_\delta)$, consider the model in Theorem 2.8. If there exists $\sqrt{n}$-consistent estimator of β for all $\vartheta \in \Theta(s_\gamma, s_\delta)$, then* $\max(s_\gamma, s_\delta) = \mathcal{O}(\sqrt{n}/\log(p))$.

3. Inference for σ_μ^2 and σ_ε^2

In this section, we consider the problem of conducting inference for both σ_μ^2 and σ_ε^2. Dicker [8], Janson, Barber and Candès [14], and Cai and Guo [5] provide interesting applications of both estimation and inference to which we refer the interested reader. The main model that we consider is slightly different than that considered in the previous section. Since we are not interested in the contribution of any particular covariate, we do not need to distinguish X from Z. Hence, we will set $q = 0$ and consider the following model,

$$Y = \mu + \varepsilon. \tag{3.1}$$

Unlike Section 2, we view μ as a random quantity, with $\sigma_\mu^2 \triangleq \mathrm{Var}(\mu_1)$. Thus, σ_μ^2 can be viewed as the explained variation in the data using the covariates Z. Throughout this section, S_γ will denote the weakly sparse set for μ with sparsity s_γ. When constructing a $\sqrt{n}$-consistent estimator for σ_μ^2, the asymptotic distribution will depend on the variance of μ_1^2, which we will denote by $\kappa_\mu \triangleq \mathrm{Var}(\mu_1^2)$. Similarly, we will need to let $\kappa_\varepsilon \triangleq \mathrm{Var}(\varepsilon_1^2)$ when constructing confidence intervals for σ_ε^2.

3.1. Inference for σ_μ^2

To motivate our high-dimensional procedure, we will start by considering the low-dimensional setting. Letting S_γ denote a weakly sparse set for μ relative to Z and identifying γ with a vector in $\mathbb{R}^{s_\gamma}$, we will temporarily consider the linear model

$$Y = Z_{S_\gamma}\gamma + \varepsilon. \tag{3.2}$$

The natural estimator for σ_μ^2 is given by $n^{-1}\|P_{S_\gamma}Y\|^2$. The following proposition shows that this natural estimator is in fact efficient for estimating σ_μ^2 with Gaussian errors.

Proposition 3.1. *Consider the model given in equation* (3.2). *Assume that the design Z_{S_γ} has full column rank and $s_\gamma < n$ is fixed. Then, the estimator $n^{-1}\|P_{S_\gamma}Y\|^2$ is efficient for estimating σ_μ^2.*

From the central limit theorem, it is immediate that

$$\sqrt{n}\big(n^{-1}\|P_{S_\gamma}Y\|^2 - \sigma_\mu^2\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \kappa_\mu + 4\sigma_\mu^2\sigma_\varepsilon^2\big).$$

In the high-dimensional setting, there are three natural extensions of this low-dimensional efficient estimator using exponential weighting. The first idea is to view $P_{S_\gamma}Y$ as the predicted values of Y and directly use take the squared norm of the predicted values given by exponential weighting. For $m \in \mathcal{M}_{u_\gamma}$, let $\hat{\gamma}_m$ denote the least-squares estimator for γ using the covariates Z_m and set

$$\hat{\mu} \triangleq \sum_{m\in\mathcal{M}_{u_\gamma}} w_{m,Y} Z_m \hat{\gamma}_m,$$

where $w_{m,Y}$ is defined in Section 2.1. Then, we may consider the estimator

$$\hat{\sigma}_{\mu,\mathrm{I}}^2 \triangleq \frac{1}{n}\|\hat{\mu}\|^2.$$

Alternatively, we may take the perspective that exponential weights concentrate well around the models with high predictive capacity, which would suggest aggregating the squared norms,

$$\hat{\sigma}^2_{\mu,\mathrm{II}} \triangleq \frac{1}{n} \sum_{m \in \mathcal{M}_{u_\gamma}} w_{m,Y} \| P_m Y \|^2.$$

The last estimator that we consider is inspired by the low-dimensional maximum likelihood estimator for σ^2_ε and the fact that $\mathrm{Var}(Y_1) = \sigma^2_\mu + \sigma^2_\varepsilon$:

$$\hat{\sigma}^2_{\mu,\mathrm{III}} \triangleq \frac{1}{n}\big(\|Y\|^2 - \|Y - \hat{\mu}\|^2\big).$$

Before stating the main results for these estimators, we will first provide all of our assumptions.

(C1) The mean vector μ has independent and identically distributed entries with finite fourth moment.

(C2) The entries of ε are independent of Z. Moreover, the entries of ε are independent and identically distributed sub-Gaussians with parameter K_ε.

(C3) The vector μ is weakly sparse relative to Z with sparsity s_γ. Furthermore, it is assumed that the statistician knows a sequence u_γ with $u_\gamma \geq s_\gamma$ for n sufficiently large and $u_\gamma = o(\sqrt{n}/\log(p))$.

Assumption (C1) implies that $\|\mu\|^2 = \mathcal{O}_{\mathbb{P}}(n)$. By Jensen's inequality, it is immediate that $\hat{\sigma}^2_{\mu,\mathrm{I}} \leq \hat{\sigma}^2_{\mu,\mathrm{II}} \leq \hat{\sigma}^2_{\mu,\mathrm{III}}$. However, it turns out that, under the above assumptions, these estimators are asymptotically equivalent at the $\sqrt{n}$-rate. Recall that $\kappa_\mu \triangleq \mathrm{Var}(\mu_1^2)$. The following theorem provides the asymptotic distribution of the three estimators.

Theorem 3.2. *Consider the model given in equation* (3.1). *Suppose that* $\sigma^2_\mu > 0$. *Under assumptions* (C1)–(C3),

$$\sqrt{n}\big(\hat{\sigma}^2_\mu - \sigma^2_\mu\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \kappa_\mu + 4\sigma^2_\varepsilon \sigma^2_\mu\big),$$

where $\hat{\sigma}^2_\mu$ *is either* $\hat{\sigma}^2_{\mu,\mathrm{I}}$, $\hat{\sigma}^2_{\mu,\mathrm{II}}$, *or* $\hat{\sigma}^2_{\mu,\mathrm{III}}$.

Since our interest is mainly asymptotic, we will write $\hat{\sigma}^2_\mu$ to denote generically one of the estimators for σ^2_μ. To construct confidence intervals for σ^2_μ, we will need to estimate κ_μ, which may be accomplished by considering

$$\hat{\kappa}_\mu \triangleq \frac{1}{n} \sum_{j=1}^{n} \big(\hat{\mu}_j^2 - \hat{\sigma}^2_\mu\big)^2.$$

The following proposition shows that $\hat{\kappa}_\mu$ is a consistent estimator for κ_μ.

Proposition 3.3. *Consider the model given in equation* (3.1). *Under assumptions* (C1)–(C3), $\hat{\kappa}_\mu \xrightarrow{\mathbb{P}} \kappa_\mu$.

Therefore, an asymptotic $(1-\alpha)$ confidence interval for σ^2_μ is given by

$$\left(\hat{\sigma}^2_\mu - z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\mu + 4\hat{\sigma}^2_\varepsilon \hat{\sigma}^2_\mu}{n}}, \hat{\sigma}^2_\mu + z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\mu + 4\hat{\sigma}^2_\varepsilon \hat{\sigma}^2_\mu}{n}}\right). \tag{3.3}$$

3.2. Inference for σ_ε^2

In this section, we are interested in constructing confidence intervals for σ_ε^2. In the low-dimensional setting with Gaussian errors, an estimator for σ_ε^2 is given by maximum likelihood, which may be written as

$$\hat{\sigma}^2_{\varepsilon,\mathrm{ML}} = \frac{1}{n}\|Y - P_{S_\gamma} Y\|^2.$$

From classical parametric theory, $\hat{\sigma}^2_{\varepsilon,\mathrm{ML}}$ is an efficient estimator for σ_ε^2 that achieves the information bound. A natural extension in the high-dimensional setting is to view $P_{S_\gamma} Y$ as the predicted value and consider the estimator

$$\hat{\sigma}^2_{\varepsilon,\mathrm{I}} \triangleq \frac{1}{n}\|Y - \hat{\mu}\|^2,$$

where $\hat{\mu}$ is defined in Section 3.1. Recalling that $\mathrm{Var}(Y_1) = \sigma_\mu^2 + \sigma_\varepsilon^2$, we may consider two more estimators of σ_ε^2 in light of the results of Section 3.1, which are

1.
$$\hat{\sigma}^2_{\varepsilon,\mathrm{II}} \triangleq \frac{1}{n}\|Y\|^2 - \hat{\sigma}^2_{\mu,\mathrm{II}}.$$

2.
$$\hat{\sigma}^2_{\varepsilon,\mathrm{III}} \triangleq \frac{1}{n}\|Y\|^2 - \hat{\sigma}^2_{\mu,\mathrm{I}}.$$

Again, by Jensen's inequality, it is immediate that $\hat{\sigma}^2_{\varepsilon,\mathrm{I}} \le \hat{\sigma}^2_{\varepsilon,\mathrm{II}} \le \hat{\sigma}^2_{\varepsilon,\mathrm{III}}$. Similar to before, these three estimators are asymptotically equivalent at the $\sqrt{n}$-rate and the following theorem provides the asymptotic distribution for all three.

Theorem 3.4. *Consider the model given in* (3.1) *with* $\sigma_\mu^2 > 0$. *Under assumptions* (C1)–(C3), $\sqrt{n}(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \kappa_\varepsilon)$, *where* $\hat{\sigma}_\varepsilon^2$ *is one of* $\hat{\sigma}^2_{\varepsilon,\mathrm{I}}$, $\hat{\sigma}^2_{\varepsilon,\mathrm{II}}$, *or* $\hat{\sigma}^2_{\varepsilon,\mathrm{III}}$.

This gives us an immediate corollary to estimating $\bar{d}$ from Section 2.2, which requires the following assumption:

(C2*) The vector $\varepsilon \sim \mathcal{N}_n(0, \Sigma_\varepsilon)$ is independent of Z with $\|\Sigma_\varepsilon\| = \mathcal{O}(1)$ and $\mathrm{tr}(\Sigma_\varepsilon)/n \to \bar{d} > 0$.

Corollary 3.5. *Consider the model given in equation* (3.1). *Under assumptions* (C1), (C2*), *and* (C3), $\hat{\sigma}^2_{\varepsilon,\mathrm{I}} \xrightarrow{\mathbb{P}} \bar{d}$.

Remark 3.1. Currently, in this section, we have assumed that $q = 0$ but the theory for all three estimators of σ_ε^2 are still valid when $q > 0$. In this setting, $X\beta + \mu$ is weakly sparse relative to (X, Z) with sparsity s_γ at rate $\sqrt{n}$. Therefore, by using exponential weighting with the design (X, Z), the above theorem implies that all three estimators are consistent for σ_ε^2.

Remark 3.2. In practice, one may consider a version of the three estimators dividing by $n - u_\gamma$ instead of n, consistent with the low-dimensional unbiased mean squared error estimator. Asymptotically, since $u_\gamma = o(\sqrt{n})$, they will have the same asymptotic distribution but seem to have better performance empirically in finite sample.

Again, since $\hat{\sigma}^2_{\varepsilon,\mathrm{I}}$, $\hat{\sigma}^2_{\varepsilon,\mathrm{II}}$, and $\hat{\sigma}^2_{\varepsilon,\mathrm{III}}$ are asymptotically equivalent, we will write $\hat{\sigma}^2_\varepsilon$ to denote a generically any of the three estimators. To construct confidence intervals for σ^2_ε, we will need to estimate κ_ε. The estimator that we propose is similar to $\hat{\kappa}_\mu$, namely we will defined $\hat{\kappa}_\varepsilon$ as

$$\hat{\kappa}_\varepsilon \triangleq \frac{1}{n}\sum_{j=1}^{n}\left((y_j - \hat{\mu}_j)^2 - \hat{\sigma}^2_\varepsilon\right)^2.$$

Analogous to Proposition 3.3, the following provides the consistency of $\hat{\kappa}_\varepsilon$.

Proposition 3.6. *Consider the model given in equation* (3.1). *Under assumptions* (C1)–(C3), $\hat{\kappa}_\varepsilon \xrightarrow{\mathbb{P}} \kappa_\varepsilon$.

Therefore, an asymptotic $(1-\alpha)$ confidence interval for σ^2_ε is given by

$$\left(\hat{\sigma}^2_\varepsilon - z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\varepsilon}{n}}, \hat{\sigma}^2_\varepsilon + z_{\alpha/2}\sqrt{\frac{\hat{\kappa}_\varepsilon}{n}}\right). \tag{3.4}$$

4. Implementation

In this section, we describe a method to approximate all of the proposed estimators. Since all of our estimators are based on exponential weighting, we will only detail the task of estimating $\hat{\theta}_{\mathrm{EW}}$, with the others being analogous. Then, the goal of approximating $\hat{\theta}_{\mathrm{EW}}$ can be split into the following two tasks:

1. Determining the values of the tuning parameters α_Y and u_θ.
2. Aggregating over $\binom{p}{u_\theta}$ models.

We will start with the second task. Suppose temporarily that values of α_Y and u_θ have been selected. To aggregate the models, we will follow the Metropolis Hastings scheme of Rigollet and Tsybakov [23]. Our approach slightly differs from theirs since we restrict our attention to u_θ-sparse models whereas they consider models of varying sizes.

Conditional on the data, the values of $\hat{\theta}_{\mathrm{EW}}$ and $\hat{\theta}_m$ for each $m \in \mathcal{M}_{u_\theta}$ are fixed. We may view $\mathcal{M}_{u_\theta}$ as the vertices of the Johnson graph $J(p, u_\theta, u_\theta - 1)$ (cf. Godsil and Royle [11]). Then, for each $m \in \mathcal{M}_{u_\theta}$, by assigning weight $w_{m,Y}$ to vertex m, the target $\hat{\theta}_{\mathrm{EW}}$ may be viewed as the expectation of the fixed estimators $\hat{\theta}_m$ over the graph $J(p, u_\theta, u_\theta - 1)$, conditional on the observed data. Hence, by taking a random walk over $J(p, u_\theta, u_\theta - 1)$, we may approximate $\hat{\theta}_{\mathrm{EW}}$.

Before describing the algorithm, we need to introduce a bit of notation. For any model $m \in \mathcal{M}_{u_\theta}$, we will let $\mathcal{K}_m$ denote the neighbors of m, which is given by

$$\mathcal{K}_m \triangleq \left\{k \in \mathcal{M}_{u_\theta} : |k \cap m| = u_\theta - 1\right\}.$$

Moreover, we will write $\mathrm{RSS}_m \triangleq \|P^\perp_m Y\|^2$, the residual sum of squares. Furthermore, recall that if $Z^\mathsf{T}_m Z_m$ is rank deficient, then $(Z^\mathsf{T}_m Z_m)^{-1}$ will denote any generalized inverse. Finally, let T_0 denote some burn-in time for the Markov chain and T denote the number of samples from the Markov chain. This will yield Algorithm1, which closely parallels Rigollet and Tsybakov [23].

Algorithm 1: Exponential weighting

Result: Approximates $\hat{\theta}_{\mathrm{EW}}$
Initialize a random point $m_0 \in \mathcal{M}_{u_\theta}$ and compute RSS_{m_0};
for $t = 0, \ldots, T$ **do**
  Uniformly select $k \in \mathcal{K}_{m_t}$ and compute RSS_k;
  Generate a random variable m_{t+1} by

$$m_{t+1} = \begin{cases} m_t & \text{with probability } \exp\left(-\frac{1}{\alpha_Y}(\mathrm{RSS}_k - \mathrm{RSS}_{m_t})\right); \\ k & \text{with probability } 1 - \exp\left(-\frac{1}{\alpha_Y}(\mathrm{RSS}_k - \mathrm{RSS}_{m_t})\right); \end{cases}$$

  if $t > T_0$ **then**
    Compute $\hat{\theta}_{t+1} \leftarrow (Z_{m_{t+1}}^{\mathsf{T}} Z_{m_{t+1}})^{-1} Z_{m_{t+1}}^{\mathsf{T}} Y$, embedded as a vector in $\mathbb{R}^p$;
  end
end
return

$$\frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \hat{\theta}_{t+1};$$

Then, analogous to Theorem 7.1 of Rigollet and Tsybakov [23], it will follow that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \hat{\theta}_{t+1} = \hat{\theta}_{\mathrm{EW}}, \quad \mathbb{P} \text{ almost surely.}$$

Now, for the first task, we may construct a grid of parameter points and use cross-validation to jointly tune the parameters using the above algorithm. Since both α_Y and u_θ do not need to be known exactly, but need to be tuned to be larger than a threshold, the grid can be quite coarse to ease the computational burden.

Computation in the ultrahigh-dimension is inherently difficult. In view of Zhang, Wainwright and Jordan [29], there is no polynomial time algorithm that achieves the minimax rate for prediction without the restricted eigenvalue condition. However, we do not know any algorithm that verifies the restricted eigenvalue condition in polynomial time (cf. Raskutti, Wainwright and Yu [21]). In this paper, we completely avoid assuming a condition like the restricted eigenvalue condition and therefore we cannot guarantee polynomial time convergence. Yet, the algorithm behaves well in practice, as can be seen from the simulations in the following section.

5. Simulations

We divide this section into two parts, corresponding to simulations for β and simulations for variance components σ_μ^2 and σ_ε^2. Additional simulation tables are included in the Supplement.

5.1. Simulations for β

For ease of comparison, our simulations will be similar to those given in van de Geer *et al.* [27]. For the linear models

$$Y = X\beta + \mu + \varepsilon,$$
$$X_j = N_j + H_j,$$

we will consider the setting where $n = 100$ and $p = 500$. There are a few parameters with which we will experiment: q, β, the distribution of the design and errors, the sparsities, and the signal to noise ratio. For each parameter pairing, we run 500 simulations. All confidence intervals will be constructed at the nominal 95% level.

Since the number of parameters of interest is fixed and low-dimensional, we will consider the settings where $q \in \{1, 3\}$. To assess both the coverage and the power, we will let β be a vector in $\mathbb{R}^q$ with values in $\{0, 1\}$. To experiment with the robustness to the sub-Gaussianity assumption, we will use Gaussian, double exponential, and $t(3)$ distributions for the errors, all scaled to have mean zero and unit variance. We will denote these distributions by z, e, and t, respectively. Therefore, $\sigma_\varepsilon^2 = 1$ throughout this section. The design will have the same distribution as the error, but with an equi-correlation covariance matrix. That is, we consider the covariance matrix, $\Sigma(Z)$ to be

$$\Sigma(Z)_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ \rho & \text{if } i \neq j \end{cases}$$

for $\rho \in \{0, 0.8\}$. When $q = 3$, the covariance matrix for $H^{(1)}$, denoted by $\Sigma(H)$, will also be equi-correlation,

$$\Sigma(H) = \begin{cases} \sigma_\eta^2 & \text{if } i = j, \\ 0.5\sigma_\eta^2 & \text{if } i \neq j, \end{cases}$$

where σ_η^2 is chosen so that $\operatorname{Var}(X_1) = 1$.

Similar to van de Geer *et al.* [27], we will let the sparsity $s_\gamma \in \{3, 15\}$, and, for simplicity, set $s_\delta = s_\gamma$. We will set the signal to noise ratio of μ to ε, which is given by $\sigma_\mu^2/\sigma_\varepsilon^2$, to be 2. Since large values of the signal to noise ratio (SNR) of N_j to H_j correspond to highly correlated designs, we will also consider $\mathrm{SNR}_X \triangleq \sigma_\nu^2/\sigma_\eta^2 \in \{2, 1000\}$.

For our simulations, we will say μ is weakly sparse relative to Z with sparsity s_γ at rate $\sqrt{n}$ if there exists an s_γ-sparse set S and vector γ_S such that $\operatorname{Var}(\mu_1 - (Z_S\gamma_S)_1) \leq n^{-1/2}$. In particular, we will consider vectors γ of the form

$$\gamma_j \propto \pi(j)^{-\kappa}, \quad j = 1, \ldots, p$$

for some value $\kappa > 0$ and permutation $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$. A similar approach is applied for Δ.

We will compare our estimators with a few other procedures:

1. (LS) Oracle least-squares that knows the true weakly sparse set S_γ.
2. (DLA) De-biased lasso from Dezeure *et al.* [7] as implemented in the R package `hdi`. We only apply this when $q = 1$.
3. (SILM) Simultaneous inference for high-dimensional linear models of Zhang and Cheng [28] as implemented in the R package `SILM`.

4. (DML) Double/de-biased machine learning of Chernozhukov *et al.* [6] with 4 folds using the scaled lasso of Sun and Zhang [24] as the estimation procedure as implemented in the R package `scalreg`. We only apply this when $q = 1$.
5. ($\mathrm{EW_I}$), ($\mathrm{EW_{II}}$), ($\mathrm{EW_{III}}$) Exponential weights using $\hat{\sigma}^2_{\varepsilon,\mathrm{I}}$, $\hat{\sigma}^2_{\varepsilon,\mathrm{II}}$, and $\hat{\sigma}^2_{\varepsilon,\mathrm{III}}$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

To evaluate the procedures, we use the following two measures

1. (AvgCov) Average coverage: The percentage of time the true value of β falls inside the confidence region.
2. (AvgLen) Average length: The average length of the confidence interval (only when $q = 1$).

The results are given in Table 1 and Tables S1–S11 from the Supplement. In the $q = 1$ setting with $\mathrm{SNR}_X = 2$, the coverage is comparable amongst all of the estimators. However, the de-biased lasso and the SILM procedure are slightly preferable in this regime since the length of the intervals are slightly shorter. When $\beta = 0$, $\mathrm{SNR}_X = 1000$, and $\rho = 0.8$, the coverage of the de-biased lasso is quite poor, with less than a 25% coverage against a nominal rate of 95%. The result should not be surprising since this corresponds to a setting of high correlation in the design, which weakens the compatibility condition. The double/de-biased machine learning approach has strong nominal coverage in this regime (about 100%), but the length of the intervals are significantly longer than the other procedures (about four to five times longer than exponential weighting). When $\beta = 1$, $\mathrm{SNR}_X = 1000$, and $\rho = 0.8$, we note that the SILM procedure no longer maintains nominal coverage. At first glance, it may seem odd that the oracle procedure based on least-squares does not always achieve the nominal coverage, but this is a consequence of weak sparsity. Since there is non-negligible bias in the model approximation in finite sample, this affects the empirical coverage of the oracle procedure. The results remain the same when we consider $q = 3$ and different distributions for the design and the errors. These results suggest that the compatibility assumption is crucial to the success of the lasso based procedures, and in the absence of such an assumption, the procedures based on exponential weighting maintain competitive coverage and length.

Table 1. Simulations for β with Gaussian design and errors when $q = 1$ and $\beta = 0$

	snr_X	2	2	2	2	1000	1000	1000	1000
	ρ	0	0	0.8	0.8	0	0	0.8	0.8
	s_δ, s_γ	3	15	3	15	3	15	3	15
	LS	0.946	0.880	0.946	0.958	0.942	0.908	0.938	0.930
	DLA	0.958	0.884	0.976	0.978	0.954	0.870	0.218	0.170
	SILM	0.970	0.872	0.962	0.970	0.958	0.812	0.900	0.902
AvgCov	DML	0.966	0.850	0.956	0.946	0.982	0.844	1.000	1.000
	$\mathrm{EW_I}$	0.956	0.868	0.956	0.962	0.960	0.828	0.954	0.968
	$\mathrm{EW_{II}}$	0.978	0.912	0.976	0.980	0.972	0.898	0.966	0.984
	$\mathrm{EW_{III}}$	0.984	0.938	0.984	0.994	0.980	0.936	0.980	0.994
	LS	0.427	0.462	0.589	0.684	0.430	0.467	0.919	1.440
	DLA	0.493	0.532	0.689	0.700	0.530	0.547	0.544	0.501
	SILM	0.529	0.559	0.670	0.697	0.623	0.609	0.666	0.646
AvgLen	DML	0.650	0.634	0.694	0.692	1.510	0.881	10.600	11.100
	$\mathrm{EW_I}$	0.623	0.636	0.700	0.716	1.060	0.774	1.910	1.830
	$\mathrm{EW_{II}}$	0.690	0.710	0.768	0.797	1.170	0.868	2.100	2.040
	$\mathrm{EW_{III}}$	0.749	0.776	0.830	0.871	1.280	0.951	2.270	2.240

5.2. Simulations for σ_μ^2 and σ_ε^2

In this section, we set $q = 0$ and only consider the setting of strong sparsity (ie. $\mu = Z\gamma$ for some vector $\gamma \in \mathbb{R}^p$ satisfying $\|\gamma\|_0 = s_\gamma$). This reduces the linear model to $Y = Z\gamma + \varepsilon$. We still consider the setting where $n = 100$ and $p = 500$. The value of $\sigma_\mu^2 = 2$ and $\sigma_\varepsilon^2 = 1$ throughout these simulations. The parameters with which we will experiment are the distributions of the design and errors and the sparsity.

Again, we will consider Gaussian, double exponential, and $t(3)$ distributions for the design and the errors. The design will have an equi-correlation structure with $\rho \in \{0, 0.8\}$ and the sparsity will satisfy $s_\gamma \in \{3, 15\}$.

The vector of coefficients, γ, will have s_γ components generated from uniform$(-1, 1)$ and $p - s_\gamma$ components that are zero. The values will then be scaled such that $\sigma_\mu^2 = \gamma^\top \Sigma_Z \gamma = 2$.

For estimation of σ_μ^2, we will compare our results with an oracular estimator based on low-dimensional least-squares and the recent proposal of CHIVE.

1. (LS) Oracle least-squares that knows the true strongly sparse set S_γ using equation (3.3).
2. (CHIVE) The calibrated inference for high-dimensional variance explained of Cai and Guo [5]. We follow Algorithm 1 of the paper with $\tau_0^2 \in \{0, 2, 4, 6\}$.
3. (EW_I), (EW_II), (EW_III) Exponential weighting using $\hat{\sigma}^2_{\mu,\text{I}}$, $\hat{\sigma}^2_{\mu,\text{II}}$, and $\hat{\sigma}^2_{\mu,\text{III}}$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

The results are presented in Table 2 and Table S12 from the Supplement. We note that the coverage of the least-squares procedure is close to the nominal 95% rate when $s_\gamma = 3$ and the errors are either Gaussian or double exponential. The coverage is significantly worse for the $t(3)$ design, which should not be surprising since the fourth moment is not defined for this distribution. However, when $s_\gamma = 15$, the coverage of least-squares falls, which establishes a reference for the problem difficulty, since Proposition 3.1 establishes the efficiency of least-squares in this problem.

Table 2. Simulations for σ_μ^2 with $s_\gamma = 3$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.922	0.948	0.914	0.934	0.808	0.802
	CHIVE_0	0.698	0.532	0.690	0.604	0.554	0.526
	CHIVE_2	0.818	0.668	0.792	0.702	0.712	0.634
	CHIVE_4	0.888	0.748	0.848	0.762	0.770	0.704
	CHIVE_6	0.890	0.772	0.898	0.790	0.860	0.746
	EW_I	0.852	0.850	0.854	0.862	0.780	0.778
	EW_II	0.804	0.772	0.820	0.838	0.812	0.828
	EW_III	0.708	0.644	0.744	0.762	0.820	0.866
AvgLen	LS	1.520	1.510	1.800	1.950	2.430	2.950
	CHIVE_0	0.998	0.937	1.160	1.190	1.670	2.130
	CHIVE_2	1.520	1.560	1.650	1.740	2.150	2.640
	CHIVE_4	1.890	1.970	2.010	2.120	2.500	2.980
	CHIVE_6	2.210	2.300	2.310	2.440	2.780	3.270
	EW_I	1.470	1.440	1.750	1.850	2.390	2.840
	EW_II	1.420	1.390	1.710	1.810	2.370	2.810
	EW_III	1.370	1.320	1.670	1.760	2.340	2.780

Amongst the exponential weighting estimators, when $s_\gamma = 3$ and the errors are Gaussian or double exponential, the procedure based on $\hat{\sigma}^2_{\mu,\mathrm{I}}$ has the best performance and $\hat{\sigma}^2_{\mu,\mathrm{III}}$ has the coverage when the errors are t distributed. For higher sparsity, no one estimators dominates the others; depending on our assumptions, any of the three estimators may be preferable. Compared with CHIVE, the best exponential weighting procedure seems to be able to achieve comparable coverage with significantly shorter intervals, which can be seen across all of our simulation settings.

For the estimation of σ^2_ε, we will consider the oracular least-squares, the scaled lasso estimator, and the refitted cross-validation with Sure Independence Screening, along with our proposed procedures based on exponential weighting.

1. (LS) Oracle least-squares that knows the true strongly sparse set S_γ using equation (3.4).
2. (SL) Scaled lasso as implemented in the R package `scalreg` with a confidence interval constructed using Theorem 2 of Sun and Zhang [24].
3. (RCV-SIS) Refitted cross-validation of Fan, Guo and Hao [9] using the Sure Independence Screening of Fan and Lv [10] as implemented in the R package `SIS` in the first stage. The confidence interval is constructed using Theorem 2 of Fan, Guo and Hao [9], with $\mathbb{E}\varepsilon^4$ estimated by Proposition 3.6 of the present paper.
4. ($\mathrm{EW_I}$), ($\mathrm{EW_{II}}$), ($\mathrm{EW_{III}}$) Exponential weighting using $\hat{\sigma}^2_{\varepsilon,\mathrm{I}}$, $\hat{\sigma}^2_{\varepsilon,\mathrm{II}}$, and $\hat{\sigma}^2_{\varepsilon,\mathrm{III}}$ respectively. We tune the parameters using cross-validation with $T_0 = 3000$ and $T = 7000$.

The results are given in Table 3 and Table S13 from the Supplement. When the signal is very sparse, $s_\gamma = 3$, and there is no correlation in the design, scaled lasso has better coverage than exponential weighting. However, as the correlation increases to $\rho = 0.8$, the confidence intervals constructed using $\hat{\sigma}^2_{\varepsilon,\mathrm{II}}$ outperforms scaled lasso both in terms of coverage and average length. When the model is less sparse, $\hat{\sigma}^2_{\varepsilon,\mathrm{I}}$ has comparable or better performance than scaled lasso. The poor performance of refitted cross-validation with Sure Independence Screening in the $s_\gamma = 15$ case should not come as a surprise since the signal to noise ratio is kept constant. The task of sure screening 15 active covariates out of 500 with low signal strength from 50 observations is very difficult.

Table 3. Simulations for σ^2_ε with $s_\gamma = 3$

	Distribution	z	z	e	e	t	t
	ρ	0	0.8	0	0.8	0	0.8
AvgCov	LS	0.938	0.912	0.952	0.940	0.918	0.912
	SL	1.000	0.730	0.998	0.730	0.994	0.756
	RCV-SIS	0.684	0.646	0.688	0.644	0.638	0.606
	$\mathrm{EW_I}$	0.616	0.608	0.678	0.674	0.650	0.690
	$\mathrm{EW_{II}}$	0.862	0.828	0.872	0.846	0.852	0.814
	$\mathrm{EW_{III}}$	0.672	0.458	0.660	0.488	0.636	0.430
AvgLen	LS	0.532	0.529	0.545	0.528	0.534	0.534
	SL	0.599	0.670	0.602	0.665	0.602	0.659
	RCV-SIS	0.485	0.509	0.508	0.514	0.554	0.539
	$\mathrm{EW_I}$	0.430	0.427	0.442	0.438	0.435	0.447
	$\mathrm{EW_{II}}$	0.441	0.444	0.453	0.453	0.446	0.463
	$\mathrm{EW_{III}}$	0.462	0.475	0.473	0.480	0.466	0.492

6. Proofs

6.1. Proofs for Section 2.1

Before proving our main results, we will state a simplified version of Theorem 2.1 of Hsu, Kakade and Zhang [13] will be useful in the subsequent proofs.

Lemma 6.1 (Theorem 2.1 of Hsu, Kakade and Zhang [13]). *Let $P \in \mathbb{R}^{n\times n}$ be a rank u projection matrix. Let $\xi \in \mathbb{R}^n$ be a mean zero sub-Gaussian vector with parameter K_ξ. Then, for all $t > 0$,*

$$\mathbb{P}\big(\|P\xi\|^2 > K_\xi^2(u + 2\sqrt{ut} + 2t)\big) \le \exp(-t).$$

For ease of reference in later proofs, we will prove Proposition 2.1 as two lemmata.

Lemma 6.2. *Let $\{w_m : w_m \ge 0, \sum_{m\in\mathcal{M}_u} w_m = 1, m \in \mathcal{M}_u\}$ be weights, possibly random, and ξ be a sub-Gaussian vector with parameter K_ξ, independent of Z. If $u = o(n^\tau/\log(p))$, then*

$$\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m \|P_m\xi\|^2\bigg) = o\big(n^\tau\big).$$

Proof. Fix $t > 0$ arbitrarily. Define the event $\mathscr{T}_t$ as

$$\mathscr{T}_t \triangleq \bigcap_{m\in\mathcal{M}_u} \big\{\|P_m\xi\|^2 \le K_\xi^2\big(u + 2\sqrt{utn^\tau} + 2tn^\tau\big)\big\}.$$

For any fixed $m \in \mathcal{M}_u$, it follows from Lemma 6.1 that

$$\mathbb{P}\big(\|P_m\xi\|^2 > K_\xi^2\big(u + 2\sqrt{utn^\tau} + 2tn^\tau\big)\big) \le \exp\big(-tn^\tau\big).$$

Therefore,

$$\mathbb{P}\big(\mathscr{T}_t^{\mathsf{c}}\big) \le \exp\big(-tn^\tau + \log\big(|\mathcal{M}_u|\big)\big). \tag{6.1}$$

We observe that the above tends to zero from the assumption that $u\log(p) = o(n^\tau)$ and the standard bound on binomial coefficients $|\mathcal{M}_u| = \binom{p}{u} \le (ep/u)^u$. Now, note that

$$\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2\bigg) = \mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2\mathbb{1}_{\mathscr{T}_t}\bigg) + \mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2\mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\bigg).$$

For the first term, by the definition of $\mathscr{T}_t$,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2\mathbb{1}_{\mathscr{T}_t}\bigg) \le 2tK_\xi^2.$$

For the second term, by Cauchy–Schwarz and equation (6.1), it follows that

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2\mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\bigg) \le \limsup_{n\to\infty} n^{-\tau}\mathbb{E}\big(\|\xi\|^2\mathbb{1}_{\mathscr{T}_t^{\mathsf{c}}}\big)$$

$$\leq \limsup_{n\to\infty} n^{-\tau}\mathbb{E}\big(\|\xi\|^4\big)^{1/2}\mathbb{P}\big(\mathscr{T}_t^{\mathsf{c}}\big)^{1/2}$$
$$= 0.$$

Therefore,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m \|P_m\xi\|^2\bigg) \leq 2tK_\xi^2.$$

Since $t > 0$ was arbitrary, this finishes the proof. □

Lemma 6.3. *Under the assumptions and setup of Proposition 2.1, for any sub-Gaussian vector ζ with parameter K_ζ independent of Z,*

1.
$$\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m \big\|P_m^{\perp}\mu\big\|^2\bigg) = o\big(n^\tau\big).$$

2.
$$\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m \mu^{\mathsf{T}} P_m^{\perp}\zeta\bigg) = o\big(n^\tau\big).$$

Note that ζ is not necessarily independent of ξ.

Proof. For $m \in \mathcal{M}_u$, let

$$r_m \triangleq \big\|P_m^{\perp}\mu\big\|^2.$$

Fixing $t > 0$ arbitrarily, define the set

$$\mathcal{A}_t \triangleq \big\{m \in \mathcal{M}_u : r_m \leq tn^\tau\big\}.$$

Now,

$$\mathbb{E}\bigg(\sum_{m\in\mathcal{M}_u} w_m r_m\bigg) = \mathbb{E}\bigg(\sum_{m\in\mathcal{A}_t} w_m r_m\bigg) + \mathbb{E}\bigg(\sum_{m\in\mathcal{A}_t^{\mathsf{c}}} w_m r_m\bigg).$$

By the definition of $\mathcal{A}_t$,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\bigg(\sum_{m\in\mathcal{A}_t} w_m r_m\bigg) \leq t.$$

For $\mathcal{A}_t^{\mathsf{c}}$, fix a value of $a > 0$, which will be determined later, and define the set $\mathscr{T}_a$ as

$$\mathscr{T}_a \triangleq \bigcap_{m\in\mathcal{M}_u} \big\{\|P_m\xi\|^2 \leq K_\xi^2\big(u + 2\sqrt{uan^\tau} + 2an^\tau\big)\big\}.$$

By the calculations from equation (6.1), it follows that

$$\mathbb{P}\big(\mathscr{T}_a^{\mathsf{c}}\big) \le \exp\big(-an^{\tau} + \log\big(|\mathcal{M}_u|\big)\big). \tag{6.2}$$

Moreover, note that, by assumption,

$$\limsup_{n\to\infty} \sup_{m\in\mathcal{M}_u} n^{-1} r_m \le \limsup_{n\to\infty} n^{-1}\|\mu\|^2 \le C,$$

for some constant $C > 0$. Then, for n sufficiently large,

$$n^{-\tau}\mathbb{E}\bigg(\sum_{m\in\mathcal{A}_t^{\mathsf{c}}} w_m r_m\bigg) \le 2Cn^{1-\tau} \sum_{m\in\mathcal{A}_t^{\mathsf{c}}} \mathbb{E}(w_m) \le 2Cn^{1-\tau} \sum_{m\in\mathcal{A}_t^{\mathsf{c}}} \big(\mathbb{E}(w_m \mathbb{1}_{\mathscr{T}_a}) + \mathbb{P}\big(\mathscr{T}_a^{\mathsf{c}}\big)\big). \tag{6.3}$$

Fix $m \in \mathcal{A}_t^{\mathsf{c}}$ temporarily and let S be any weakly sparse set for μ. Then, we have that

$$\begin{aligned} w_m \mathbb{1}_{\mathscr{T}_a} &\le \exp\bigg(-\frac{1}{\alpha}\big(\big\|P_m^{\perp} Y\big\|^2 - \big\|P_S^{\perp} Y\big\|^2\big)\bigg)\mathbb{1}_{\mathscr{T}_a} \\ &\le \exp\bigg(-\frac{1}{\alpha}\big(r_m - r_S + 2\mu^{\mathsf{T}} P_m^{\perp}\xi - 2\mu^{\mathsf{T}} P_S^{\perp}\xi - K_\xi^2\big(u + 2\sqrt{uan^{\tau}} + 2an^{\tau}\big)\big)\bigg). \end{aligned}$$

By Cauchy–Schwarz,

$$\begin{aligned} \mathbb{E}(w_m \mathbb{1}_{\mathscr{T}_a}) \le{}& \exp\bigg(-\frac{1}{\alpha}\big(r_m - r_S - K_\xi^2\big(u + 2\sqrt{uan^{\tau}} + 2an^{\tau}\big)\big)\bigg) \\ &\times \bigg(\mathbb{E}\exp\bigg(-\frac{4}{\alpha}\mu^{\mathsf{T}} P_m^{\perp}\xi\bigg)\bigg)^{1/2} \bigg(\mathbb{E}\exp\bigg(\frac{4}{\alpha}\mu^{\mathsf{T}} P_S^{\perp}\xi\bigg)\bigg)^{1/2}. \end{aligned}$$

Computing each of the Laplace transforms directly, it follows that

$$\mathbb{E}\exp\bigg(-\frac{4}{\alpha}\mu^{\mathsf{T}} P_m^{\perp}\xi\bigg) \le \exp\bigg(\frac{8K_\xi^2}{\alpha^2} r_m\bigg).$$

Here, we have used Definition 1.2. Similarly,

$$\mathbb{E}\exp\bigg(\frac{4}{\alpha}\mu^{\mathsf{T}} P_S^{\perp}\xi\bigg) \le \exp\bigg(\frac{8K_\xi^2}{\alpha^2} r_S\bigg).$$

Hence,

$$\begin{aligned} \mathbb{E}(w_m \mathbb{1}_{\mathscr{T}_a}) &\le \exp\bigg(-\frac{1}{\alpha}\bigg(\bigg(1 - \frac{4K_\xi^2}{\alpha}\bigg) r_m - \bigg(1 + \frac{4K_\xi^2}{\alpha}\bigg) r_S - K_\xi^2\big(u + 2\sqrt{uan^{\tau}} + 2an^{\tau}\big)\bigg)\bigg) \\ &\le \exp\bigg(-\frac{1}{\alpha}\bigg(\bigg(1 - \frac{4K_\xi^2}{\alpha}\bigg) tn^{\tau} - \bigg(1 + \frac{4K_\xi^2}{\alpha}\bigg) r_S - K_\xi^2\big(u + 2\sqrt{uan^{\tau}} + 2an^{\tau}\big)\bigg)\bigg). \end{aligned}$$

The second inequality follows from the fact that $m \in \mathcal{A}_t^{\mathsf{c}}$. Since $u = o(n^{\tau}/\log(p))$, setting $a < (1 - 4K_\xi^2/\alpha)t/2$ yields

$$\mathbb{E}(w_m \mathbb{1}_{\mathscr{T}_a}) \le \exp\bigg(-\frac{1}{\alpha}\bigg(\bigg(1 - \frac{4K_\xi^2}{\alpha}\bigg) t - 2a\bigg) n^{\tau} + o\big(n^{\tau}\big)\bigg). \tag{6.4}$$

Combining equations (6.2), (6.3), and (6.4), it follows that

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\Big(\sum_{m\in\mathcal{A}_t^{\mathsf{c}}} w_m r_m\Big) = 0.$$

Therefore,

$$\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\Big(\sum_{m\in\mathcal{M}_u} w_m r_m\Big) \le t.$$

Since $t > 0$ was arbitrary, the first claim follows. For the second half, let the set $\mathscr{F}_t$ be

$$\mathscr{F}_t \triangleq \bigcap_{m\in\mathcal{A}_t} \big\{\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big| \le tn^{\tau}\big\}.$$

For a fixed $m \in \mathcal{A}_t$, it will follow by a Chernoff bound that, for some constant $c > 0$,

$$\mathbb{P}\big(\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big| > tn^{\tau}\big) \le 2\exp\Big(-\frac{ct^2n^{2\tau}}{K_\zeta^2 r_m}\Big) \le 2\exp\Big(-\frac{ctn^{\tau}}{K_\zeta^2}\Big).$$

Therefore, an upper bound for $\mathbb{P}(\mathscr{F}_t^{\mathsf{c}})$ is given by

$$\mathbb{P}\big(\mathscr{F}_t^{\mathsf{c}}\big) \le 2\exp\Big(-\frac{ctn^{\tau}}{K_\zeta^2} + \log\big(|\mathcal{A}_t|\big)\Big). \tag{6.5}$$

Now,

$$\mathbb{E}\Big(\sum_{m\in\mathcal{A}_t} w_m\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big|\Big) = \mathbb{E}\Big(\sum_{m\in\mathcal{A}_t} w_m\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big|\mathbb{1}_{\mathscr{F}_t}\Big) + \mathbb{E}\Big(\sum_{m\in\mathcal{A}_t} w_m\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big|\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\Big).$$

By the definition of $\mathscr{F}_t$, it follows that

$$\mathbb{E}\Big(\sum_{m\in\mathcal{A}_t} w_m\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big|\mathbb{1}_{\mathscr{F}_t}\Big) \le tn^{\tau}.$$

On $\mathscr{F}_t^{\mathsf{c}}$, two applications of Cauchy–Schwarz and equation (6.5) yields

$$\begin{aligned}
&\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\Big(\sum_{m\in\mathcal{A}_t} w_m\big|\mu^{\mathsf{T}} P_m^{\perp}\zeta\big|\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\Big)\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\mathbb{E}\big(\|\zeta\|\mathbb{1}_{\mathscr{F}_t^{\mathsf{c}}}\big)\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\big(\mathbb{E}\|\zeta\|^2\big)^{1/2}\big(\mathbb{P}\big(\mathscr{F}_t^{\mathsf{c}}\big)\big)^{1/2}\\
&\quad= 0.
\end{aligned}$$

Furthermore, on $\mathcal{A}_t^{\mathsf{c}}$, by another two applications of Cauchy–Schwarz,

$$\begin{aligned}
&\limsup_{n\to\infty} n^{-\tau}\mathbb{E}\Big(\sum_{m\in\mathcal{A}_t^{\mathsf{c}}} w_m\big|\mu^{\mathsf{T}}P_m^{\perp}\zeta\big|\Big)\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\sum_{m\in\mathcal{A}_t^{\mathsf{c}}}\mathbb{E}\big(w_m\|\zeta\|\big)\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\sum_{m\in\mathcal{A}_t^{\mathsf{c}}}\big(\mathbb{E}w_m^2\big)^{1/2}\big(\mathbb{E}\|\zeta\|^2\big)^{1/2}\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\big(\mathbb{E}\|\zeta\|^2\big)^{1/2}\sum_{m\in\mathcal{A}_t^{\mathsf{c}}}\big(\mathbb{E}w_m\big)^{1/2}\\
&\quad\le \limsup_{n\to\infty} n^{-\tau}\|\mu\|\big(\mathbb{E}\|\zeta\|^2\big)^{1/2}\sum_{m\in\mathcal{A}_t^{\mathsf{c}}}\big(\mathbb{E}(w_m\mathbb{1}_{\mathscr{T}_a})+\mathbb{P}\big(\mathscr{T}_a^{\mathsf{c}}\big)\big)^{1/2}\\
&\quad= 0,
\end{aligned}$$

where the limit follows by equations (6.2) and (6.4). Since $t>0$ was arbitrary, this proves the second claim and finishes the proof. □

Immediately, we have the following corollary for random designs when the mean vector is assumed to be weakly sparse in probability.

Corollary 6.4. *Consider the setup of Lemma 6.3. If μ is weakly sparse relative to Z in probability and $\|\mu\|^2=\mathcal{O}_{\mathbb{P}}(n^{\tau})$, then*

1.
$$\Big(\sum_{m\in\mathcal{M}_u} w_m\big\|P_m^{\perp}\mu\big\|^2\Big)=o_{\mathbb{P}}\big(n^{\tau}\big).$$

2.
$$\Big(\sum_{m\in\mathcal{M}_u} w_m\mu^{\mathsf{T}}P_m^{\perp}\zeta\Big)=o_{\mathbb{P}}\big(n^{\tau}\big).$$

With these lemmata, we can now prove Proposition 2.1.

Proof of Proposition 2.1. Indeed, by convexity of the norm, it follows that

$$\Big\|\sum_{m\in\mathcal{M}_u} w_m Z\hat{\gamma}_m-\mu\Big\|^2\le\sum_{m\in\mathcal{M}_u} w_m\big\|P_m^{\perp}\mu\big\|^2+\sum_{m\in\mathcal{M}_u} w_m\|P_m\xi\|^2.$$

Applying Lemmata 6.2 and 6.3 finishes the proof. □

Instead of directly proving Theorem 2.3, we will decompose the estimator and prove each part separately. Indeed, we note that

$$\hat{\beta}_{\mathrm{EW}} = \frac{(\nu - Z\hat{\delta}_{\mathrm{EW}} + \eta)^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}} + \eta\beta + \varepsilon)}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2}.$$

Then,

$$\begin{aligned}\sqrt{n}\hat{\beta}_{\mathrm{EW}} = \big((\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}} + \eta\beta + \varepsilon) + \eta^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}) \\ + \eta^{\mathsf{T}}\eta\beta + \eta^{\mathsf{T}}\varepsilon\big) \times \frac{1}{\sqrt{n}\sigma_\eta^2} \times \frac{n\sigma_\eta^2}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2}.\end{aligned}$$

We will start by proving that the first line, which corresponds to the bias from inexact orthogonalization, converges to zero.

Lemma 6.5. *Consider the models given in equations* (2.1) *and* (2.2). *Under assumptions* (A1)–(A3),

$$(\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}} + \eta\beta + \varepsilon) + \eta^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}) = o_{\mathbb{P}}(\sqrt{n}).$$

Proof. Without the loss of generality, we will assume that $u \triangleq u_\theta = u_\delta$. Expanding, we have

$$(\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}) + (\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\eta\beta + \varepsilon) + \eta^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}).$$

We will consider each of the three terms separately. By Cauchy–Schwarz and Corollary 2.2, it follows that

$$\big|(\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}})\big| \le \|\nu - Z\hat{\delta}_{\mathrm{EW}}\|\|\mu - Z\hat{\theta}_{\mathrm{EW}}\| = o_{\mathbb{P}}(\sqrt{n}).$$

For the second term, we may further expand to obtain

$$\begin{aligned}(\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\eta\beta + \varepsilon) &= \sum_{m\in\mathcal{M}_u} w_{m,X}\big(P_m^{\perp}\nu - P_m\eta\big)^{\mathsf{T}}(\eta\beta + \varepsilon) \\ &= \sum_{m\in\mathcal{M}_u} w_{m,X}\nu^{\mathsf{T}}P_m^{\perp}(\eta\beta + \varepsilon) + \frac{1}{2}\sum_{m\in\mathcal{M}_u} w_{m,X}\|P_m\varepsilon\|^2 \\ &\quad - \frac{1}{2}\sum_{m\in\mathcal{M}_u} w_{m,X}\big\|P_m(\eta + \varepsilon)\big\|^2 \\ &\quad - \left(\beta - \frac{1}{2}\right)\sum_{m\in\mathcal{M}_u} w_{m,X}\|P_m\eta\|^2.\end{aligned}$$

In the model $X = \nu + \eta$, applying Lemma 6.2 with $\xi = \varepsilon$, $\xi = \eta + \varepsilon$, and $\xi = \eta$ and Corollary 6.4 with $\zeta = \eta\beta + \varepsilon$ will imply that

$$(\nu - Z\hat{\delta}_{\mathrm{EW}})^{\mathsf{T}}(\eta\beta + \varepsilon) = o_{\mathbb{P}}(\sqrt{n}).$$

Finally,

$$\begin{aligned}\eta^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}) &= \sum_{m\in\mathcal{M}_u} w_{m,Y}\eta^{\mathsf{T}}\big(P_m^{\perp}\mu - P_m(\eta\beta+\varepsilon)\big)\\ &= \sum_{m\in\mathcal{M}_u} w_{m,Y}\eta^{\mathsf{T}}P_m^{\perp}\mu - \frac{1}{2}\sum_{m\in\mathcal{M}_u} w_{m,Y}\big\|P_m\big(\eta(\beta+1)+\varepsilon\big)\big\|^2\\ &\quad + \frac{1}{2}\sum_{m\in\mathcal{M}_u} w_{m,Y}\big\|P_m(\eta\beta+\varepsilon)\big\|^2 + \frac{1}{2}\sum_{m\in\mathcal{M}_u} w_{m,Y}\|P_m\eta\|^2.\end{aligned}$$

To finish the proof, we similarly apply Lemma 6.2 and Corollary 6.4 in the model $Y = \mu + \eta\beta + \varepsilon$. It follows that

$$\eta^{\mathsf{T}}(\mu - Z\hat{\theta}_{\mathrm{EW}}) = o_{\mathbb{P}}(\sqrt{n}). \qquad \square$$

Lemma 6.6. *Consider the models given in* (2.1) *and* (2.2). *Under assumptions* (A1)–(A3),

1.
$$\sqrt{n}\left(\frac{\eta^{\mathsf{T}}\eta\beta}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2} - \beta\right) \xrightarrow{\mathbb{P}} 0.$$

2.
$$n^{-1/2}\frac{\eta^{\mathsf{T}}\varepsilon}{\sigma_\eta^2} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\right).$$

3.
$$\frac{n\sigma_\eta^2}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2} \xrightarrow{\mathbb{P}} 1.$$

Proof. Indeed, expanding the denominator, we see that

$$\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2 = \|\nu - Z\hat{\delta}_{\mathrm{EW}}\|^2 + 2\eta^{\mathsf{T}}(\nu - Z\hat{\delta}_{\mathrm{EW}}) + \|\eta\|^2.$$

By Corollary 2.2 and Lemma 6.5, it follows that

$$\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2 = o_{\mathbb{P}}(\sqrt{n}) + \|\eta\|^2.$$

Then, by the Law of Large Numbers, $n^{-1}\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2 \xrightarrow{\mathbb{P}} \sigma_\eta^2$. This proves the third claim. Now, by direct substitution, we have that

$$\sqrt{n}\left(\frac{(\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2 + o_{\mathbb{P}}(\sqrt{n}))\beta}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2} - \beta\right) = \frac{n}{\|X - Z\hat{\delta}_{\mathrm{EW}}\|^2}\frac{o_{\mathbb{P}}(\sqrt{n})}{\sqrt{n}} = o_{\mathbb{P}}(1),$$

which proves the first claim. The second claim follows by the Central Limit Theorem, which finishes the proof. $\square$

Proof of Theorem 2.3. The proof follows by combining Lemmata 6.5 and 6.6. $\square$

Proof of Corollary 2.4. By possibly enlarging $\mathscr{K}$, we note that $\mathscr{K}$ can be written as

$$\mathscr{K} = \big\{\big(\beta, \sigma_\eta^2, \sigma_\varepsilon^2, K_\eta, K_\varepsilon\big) : |\beta| \le \beta_U, \sigma_\eta^2 \in \big[\sigma_{\eta,L}^2, \sigma_{\eta,U}^2\big], \sigma_\varepsilon^2 \in \big[\sigma_{\varepsilon,L}^2, \sigma_{\varepsilon,U}^2\big], \\ K_\eta \in [K_{\eta,L}, K_{\eta,U}], K_\varepsilon \in [K_{\varepsilon,L}, K_{\varepsilon,U}]\big\}$$

for fixed positive constants β_U, $\sigma_{\eta,L}^2$, $\sigma_{\eta,U}^2$, $\sigma_{\varepsilon,L}^2$, $\sigma_{\varepsilon,U}^2$, $K_{\eta,L}$, $K_{\eta,U}$, $K_{\varepsilon,L}$, and $K_{\varepsilon,U}$. Observe that the vectors η, $\eta\beta$, and ε are uniformly sub-Gaussian with parameters $K_{\eta,U}$ $\beta_U K_{\eta,U}$, and $K_{\varepsilon,U}$ for $\vartheta \in \mathscr{K}$ respectively. Thus, applications of Lemmata 6.2 and 6.3 are uniform. Therefore, Lemmata 6.5 and 6.6 will also hold uniformly for $\vartheta \in \mathscr{K}$, which will prove the claim. □

6.2. Proofs for Section 2.4

Proof of Theorem 2.8. Suppose that $s_\delta = o(\sqrt{n}/\log(p))$. We will consider a sequence of $\vartheta \in \Theta(s_\gamma, s_\delta)$ such that $S_\gamma \cap S_\delta = \varnothing$ and $\delta \ge 0$ componentwise. We will construct $\Sigma_{Z,Z}$ implicitly. For $j \in S_\delta^{\mathsf{c}}$, let

$$Z_j \overset{i.i.d.}{\sim} \mathcal{N}_n(0_n, I_n).$$

Before defining Z_j for $j \in S_\delta$, we will need to define another Gaussian matrix $\Xi \in \mathbb{R}^{n\times p}$. For $j \in S_\delta^{\mathsf{c}}$, set $\Xi_j = 0_n$. Then, for $j \in S_\delta$,

$$\Xi_j \overset{i.i.d.}{\sim} \mathcal{N}_n\big(0_n, \tau_n^2 I_n\big),$$

independent of Z_k for all $k \in S_\delta^{\mathsf{c}}$; the value $\tau_n^2 > 0$ will be determined later. Now, for $j \in S_\delta$, we will let $Z_j = Z\gamma + \Xi_j$. Therefore, it follows that $Z\delta = Z\gamma\|\delta\|_1 + \Xi\delta$. By a direct calculation,

$$\operatorname{Cov}\big((Z\delta)_1, (Z\gamma)_1\big) = \operatorname{Cov}\big((Z\gamma)_1\|\delta\|_1 + (\Xi\delta)_1, (Z\gamma)_1\big) = \operatorname{Var}\big((Z\gamma)_1\big)\|\delta\|_1.$$

Moreover,

$$\operatorname{Var}\big((Z\delta)_1\big) = \operatorname{Var}\big((Z\gamma)_1\|\delta\|_1 + (\Xi\delta)_1\big) = \operatorname{Var}\big((Z\gamma)_1\big)\|\delta\|_1^2 + \tau_n^2\|\delta\|_2^2.$$

Choosing $\tau_n^2 \to 0$ sufficiently fast, it will follow that

$$\operatorname{Var}\big((Z\delta)_1\big) = \operatorname{Var}\big((Z\gamma)_1\big)\|\delta\|_1^2 + o\big(n^{-1/2}\big).$$

Hence, this implies that

$$\operatorname{Cov}\big((Z\delta)_1, (Z\gamma)_1\big) = \sqrt{\operatorname{Var}\big((Z\delta)_1\big)\operatorname{Var}\big((Z\gamma)_1\big)} + o\big(n^{-1/2}\big).$$

Now, note that

$$\operatorname{Cov}\big((Z\delta)_1, (Z\gamma)_1\big) = \operatorname{Cov}(X_1, Y_1) - \beta\operatorname{Var}(X_1).$$

Let $\hat\beta$ be any $\sqrt{n}$-consistent estimator for β. Then, $n^{-1}(X^{\mathsf{T}}Y - \hat\beta X^{\mathsf{T}}X)$ is a $\sqrt{n}$-consistent estimator for $\operatorname{Cov}((Z\delta)_1, (Z\gamma)_1)$. Consider an oracle that has access to the set S_δ, knows $S_\delta \cap S_\gamma = \varnothing$, and knows the covariance structure of the design. Then, since $s_\delta = o(\sqrt{n}/\log(p))$, a $\sqrt{n}$-consistent estimator for $\operatorname{Var}((Z\delta)_1)$ is given by Theorem 3.2. This implies that there exists a $\sqrt{n}$-consistent estimator for

$\mathrm{Var}((Z\gamma)_1)$. By the minimax lower bounds established by Cai and Guo [5], it follows that, in order to have a $\sqrt{n}$-consistent estimator for $\mathrm{Var}((Z\gamma)_1)$, it must be the case that $s_\gamma = \mathcal{O}(\sqrt{n}/\log(p))$. This proves half of the claim. The other half follows by symmetry, which finishes the proof. □

Acknowledgements

We would like to thank the anonymous referees for their helpful comments. This work was supported in part by NSF Grants DMS-1646108 and DMS-1712962.

Supplementary Material

Supplement to "Inference without compatibility: Using exponential weighting for inference on a parameter of a linear model." (DOI: 10.3150/20-BEJ1280SUPP; .pdf). In the supplement, we provide additional simulation tables along with the proofs for the remaining results.

References

[1] Bellec, P.C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. arXiv preprint arXiv:1804.01230.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[3] Bradic, J., Claeskens, G. and Gueuning, T. (2020). Fixed effects testing in high-dimensional linear mixed models. *J. Amer. Statist. Assoc.* 1–16.

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[5] Cai, T.T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 391–419. MR4084169

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[7] Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

[8] Dicker, L.H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284. MR3215347 https://doi.org/10.1093/biomet/ast065

[9] Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839 https://doi.org/10.1111/j.1467-9868.2011.01005.x

[10] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

[11] Godsil, C. and Royle, G. (2001). *Algebraic Graph Theory*. *Graduate Texts in Mathematics* **207**. New York: Springer. MR1829620 https://doi.org/10.1007/978-1-4613-0163-9

[12] Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979 https://doi.org/10.1214/17-BA1085

[13] Hsu, D., Kakade, S.M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877 https://doi.org/10.1214/ECP.v17-2079

[14] Janson, L., Barber, R.F. and Candès, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065. MR3689308 https://doi.org/10.1111/rssb.12203
[15] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[16] Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630
[17] Law, M. and Ritov, Y. (2021). Supplement to "Inference without compatibility: Using exponential weighting for inference on a parameter of a linear model." https://doi.org/10.3150/20-BEJ1280SUPP
[18] Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371
[19] Leung, G. and Barron, A.R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172
[20] Li, S., Cai, T.T. and Li, H. (2019). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. arXiv preprint arXiv:1907.06116.
[21] Raskutti, G., Wainwright, M.J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855
[22] Reid, S., Tibshirani, R. and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statist. Sinica* **26** 35–67. MR3468344
[23] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337 https://doi.org/10.1214/10-AOS854
[24] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166 https://doi.org/10.1093/biomet/ass043
[25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[26] van de Geer, S.A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506
[27] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
[28] Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. MR3671768 https://doi.org/10.1080/01621459.2016.1166114
[29] Zhang, Y., Wainwright, M.J. and Jordan, M.I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory* 921–948.
[30] Zhang, C.-H. and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

Received January 2020 and revised September 2020